

# Web数据抽取神器

- Connotate是一家为美联社、路透社、道琼斯等大型公司对全球上千个网站的非结构化数据进行实时分类和分析的公司
    - 新泽西州立大学
  - 特点
    - 自建Agent
    - 支持SOAP与REST Web Services APIs
    - ODBC-SQL、MySQL、Oracle
    - 卖点
      - 支持PDF
      - 支持图形界面
      - 支持数据库
      - 机器学习→研判结果→报告
      - 信息有效率90%？
  - 成本
    - 三台服务器
      - 数据库服务器（数据抽取的关键）
        - 2GHz双四核以上的处理器
        - 32GB以上的内存
        - 操作系统-146 GB SCSI 驱动 (RAID-1)
        - 结构化数据库&数据：450GB SCSI 驱动(RAID-10)
      - Web服务器
        - 2GHz双四核以上的处理器
        - 8GB以上的内存
        - 146 GB SCSI 驱动 (RAID-1)
      - 处理服务器
        - 2GHz双四核以上的处理器
        - 8GB以上的内存
        - 146 GB SCSI 驱动 (RAID-1)
    - Agent
      - \$40 per
  - 用途
    - Agent训练生成器
      - 训练过程视屏
        - <https://www.youtube.com/watch?v=UINBLPsa0d0>
    - 数据爬取
- 大数据厂商联盟